Arthritis & Rheumatism (Arthritis Care & Research)
Vol. 55, No. 1, February 15, 2006, pp 74–80
DOI 10.1002/art.21695

ORIGINAL ARTICLE

Short-Term Prediction of Mortality in Patients With Systemic Lupus Erythematosus: Classification of Outcomes Using Random Forests

MICHAEL M. WARD, SINISA PAJEVIC, JONATHAN DREYFUSS, AND JAMES D. MALLEY

Objective. **To identify demographic and clinical characteristics that classify patients with systemic lupus erythematosus (SLE) at risk for in-hospital mortality.**

Methods. **Patients hospitalized in California from 1996 to 2000 with a principal diagnosis of SLE (N = 3,839) were identified from a state hospitalization database. As candidate predictors of mortality, we used patient demographic characteristics; the presence or absence of 40 different clinical conditions listed among the discharge diagnoses; and 2 summary indexes derived from the discharge diagnoses, the Charlson Index and the SLE Comorbidity Index. Predictors of patients at increased risk of mortality were identified and validated using random forests, a statistical procedure that is a generalization of single classification trees. Random forests use bootstrapped samples of patients and randomly selected subsets of predictors to create individual classification trees, and this process is repeated to generate multiple trees (a forest). Classification is then done by majority vote across all trees.**

Results. **Of the 3,839 patients, 109 died during hospitalization. Selecting from all available predictors, the random forests had excellent predictive accuracy for classification of death. The mean classification error rate, averaged over 10 forests of 500 trees each, was 11.9%. The most important predictors were the Charlson Index, respiratory failure, SLE Comorbidity Index, age, sepsis, nephritis, and thrombocytopenia.**

Conclusion. **Information on clinical diagnoses can be used to accurately predict mortality among hospitalized patients with SLE. Random forests represent a useful technique to identify the most important predictors from a larger (often much larger) number and to validate the classification.**

KEY WORDS. Systemic lupus erythematosus; Mortality; Hospitalization; Classification tree; Random forest.

INTRODUCTION

Although the health outcomes of patients with systemic lupus erythematosus (SLE) have improved over the past several decades, SLE remains a potentially fatal disease (1). The availability of accurate prognostic information would allow better identification of patients at greatest risk for poor outcomes, who might then be candidates for more intensive monitoring.

Supported by the Intramural Research Program, NIH (National Institute of Arthritis and Musculoskeletal and Skin Diseases and Center for Information Technology).

Michael M. Ward, MD, MPH, Sinisa Pajevic, PhD, Jonathan Dreyfuss, BS, James D. Malley, PhD: National Institutes of Health, US Department of Health and Human Services, Bethesda, Maryland.

Address correspondence to Michael M. Ward, MD, MPH, NIH/NIAMS/IRP, Building 10 CRC, Room 4–1339, 10 Center Drive, MSC 1468, Bethesda, MD 20892–1828. E-mail: wardm1@mail.nih.gov.

Submitted for publication March 23, 2005; accepted in revised form August 18, 2005.

Studies of prognosis in SLE have focused on identifying risk factors, present either at the onset of SLE or developing during the course of illness, that predict mortality over the subsequent 1–10 years (2–24). Although results vary among studies, nephritis, thrombocytopenia, central nervous system involvement, lung disease, and summary measures of SLE activity and organ damage are the risk factors that have most consistently been found to predict mortality in patients with SLE. These studies are useful because they inform us about the ways in which specific manifestations of SLE influence health outcomes. However, to provide a prognosis for an individual patient, it is important to identify subgroups of patients at high risk for poor outcomes. Also, these studies rarely consider interactions among clinical manifestations, yet patients always present constellations of clinical features, which may interact in different ways to influence prognosis. In addition, these studies often report the risks of mortality in patients with a particular clinical manifestation compared with those without the manifestation. However, to estimate the prognosis of an individual patient, the focus should be on the absolute risk associated with their particular clinical manifestations, rather than the relative risk. Lastly, studies of long-term mortality are less helpful in guiding treatment decision making than studies of short-term prognosis. Short-term outcomes would be more likely to be influenced by timely intervention (18).

In this study, we used data from a population-based hospitalization database to characterize patients at risk for in-hospital mortality. We chose to focus on patients with active SLE, defined as patients who had an urgent or emergency hospitalization due to SLE, for whom information on short-term prognosis would be most clinically useful. To classify patients and identify predictors, we used random forests, a new method that develops numerous classification trees and uses these to test the accuracy of prediction and identify the most important predictors (25).

PATIENTS AND METHODS

Data source. Patients in this study were identified through a search of data files compiled by the California Office of Statewide Health Planning and Development (OSHPD). All acute-care, nonfederal hospitals in California are mandated to provide this agency with information on each patient discharge. The discharge abstracts include information on patient demographic characteristics; the principal diagnosis, defined as the condition chiefly responsible for the hospitalization (by International Classification of Diseases, Ninth Revision, Clinical Modification [ICD-9-CM] code) (26); up to 24 additional diagnoses; major procedures (up to 21, by ICD-9-CM codes); and disposition. The data files include ~2.6 million discharge abstracts annually. Unique patient identifiers were available, which enabled tracking of patients over time.

For this study, we abstracted data on all acute-care hospitalizations of patients with SLE (ICD-9-CM code 710.0) age ≥18 years from 1996 to 2000. The target cohort consisted of all patients with an unscheduled hospitalization (urgent or emergency admission, based on the discharge abstract, and not related to childbirth) for which the principal diagnosis was SLE. Enrollment in the cohort occurred at the first unscheduled hospitalization with a principal diagnosis of SLE that occurred in 1996–2000. Of 21,021 patients with SLE who were hospitalized during these years, 17,824 patients had at least 1 unscheduled hospitalization, and 3,839 patients had at least 1 unscheduled hospitalization for which the principal diagnosis was SLE. Interhospital transfers, which occurred in 108 patients (2.8%), were considered as a single hospitalization.

Before OSHPD data are released, the data are extensively checked for reliability and validity, and data fields with error rates ≥0.1% are returned to hospitals for correction (27,28). Reabstraction studies that have compared OSHPD data files with original medical records have found specificities for diagnoses of 0.98–1.00 and sensitivities for diagnoses of 0.88–1.00 (29–31). To protect patient confidentiality, the data are anonymous; therefore, validation of specific diagnoses in this study could not be performed.

Study variables. Prognostic variables included patient age, sex, race (white, African American, Asian, Native American, other, or unknown), Hispanic ethnicity, medical insurance status (private, Medicare, public other than Medicare, no insurance), the presence of specific manifestations of SLE (thrombocytopenia, autoimmune hemolytic anemia, pericarditis, pleuritis, nephritis, chronic renal failure, seizures, and psychosis), and common comorbid medical conditions, including those used in the Charlson Comorbidity Index and those that are common reasons for hospitalization among patients with SLE (32–34). These conditions included acute myocardial infarction, unstable angina, atrial fibrillation, congestive heart failure, chest pain not otherwise specified, cerebrovascular accident, hemiplegia, peripheral vascular disease, chronic obstructive pulmonary disease, respiratory failure, pulmonary embolus, deep venous thrombosis, peptic ulcer disease, gastritis, mild liver disease, severe liver disease, pancreatitis, renal transplantation, arteriovenous fistula complication, sepsis, pneumonia, urinary tract infection, cellulitis, osteomyelitis, acquired immunodeficiency syndrome, diabetes, cancer, metatstatic cancer, dementia, dehydration, syncope, and hip fracture or pathologic fracture. All clinical conditions used for prognosis were those recorded as one of the possible 25 discharge diagnoses for this hospitalization (by ICD-9-CM codes). We also examined the Charlson Index, a weighted sum of 16 chronic medical conditions, and the SLE Comorbidity Index, a weighted sum of 14 conditions designed to measure comorbidity in SLE, as prognostic factors (32,33).

Statistical analysis. Five demographic variables, 40 individual clinical manifestations, and 2 clinical indexes were tested as independent variables in the classification of patients as either dying or surviving during hospitalization. We were interested in prediction without regard to attribution or time of onset of the diagnosis, and the clinical predictors could have occurred at any time during the hospitalization.

We used random forests to identify and validate important predictors of mortality and to identify subsets of patients at increased risk of death (25). A random forest is an ensemble classifier that uses multiple classification trees for prediction. A single classification tree is a hierarchical classification procedure that uses recursive partitioning to identify subgroups of patients that are increasingly homogenous with respect to the outcome of interest. For example, the patient group is split into 2 subgroups based on the characteristic that best segregates patients at high risk of death from those at low risk of death, and this process is repeated for each resultant subgroup until either all patients are classified or subgroups of sufficient homogeneity are found. The procedure is nonparametric, not model based, identifies those independent variables that best segregate subgroups as important predictors, and identifies interactions among independent variables. A single classification tree prediction is commonly validated using a split-sample technique in which the tree is developed using a portion of the patients and then tested using patients not included in the development of the tree.

Random forests work by generating many classification trees, often several hundred at a minimum. Each tree is developed on a random sample of patients (a bootstrap sample, with sampling done with replacement) and using a randomly sampled subset of independent variables (sampled without replacement) at each node in the tree. The number of independent variables tested at each node is the same for all nodes in all trees, and efficiency is optimized if the number tested is approximately the square root of the number of independent variables (with 47 independent variables, we tested a random subset of 7 variables at each node). The best split at a single node is determined by the independent variable that best divides the patients in that node into 2 subgroups, each with the most pure membership possible. There are several methods available for specifying node purity; we use the Gini index throughout. This process is repeated at each subsequent node until the tree is grown to the largest extent possible. The accuracy of prediction is determined by how well each tree classifies each patient who was omitted from the development of the tree. The proportion of times that an individual test patient was misclassified by each tree (i.e., the patient died in-hospital but the tree predicted he or she would not have died, or vice versa), averaged over all patients, is considered to be a relatively unbiased estimate of classification error. This process obviates the need for a separate validation step.

The relative importance of the independent variables is determined by first counting the number of test patients correctly classified by each tree, then randomly changing the value of one independent variable of the test patients (e.g., for dichotomous variables, changing the result for nephritis from present to absent, or absent to present, randomly among the test patients; for continuous variables such as age, randomly changing the true value to another value present in the sample of patients), dropping the test patients with the permuted independent variable values down the tree, and subtracting the number of test patients with the permuted independent variable value who were correctly classified from the number of test patients who were correctly classified when the true unaltered values of the independent variable were used. This difference is averaged over all trees and repeated for each independent variable. A large difference in the number of patients correctly classified when the independent variable is intentionally altered indicates that the independent variable is important for correct classification, whereas a small difference indicates that the independent variable is less important for correct classification. Relative importance provides a measure by which predictors can be ranked with respect to each other, but is not a measure of relative risk associated with the predictor, and cannot be applied to risk estimation for individual patients. Random forests can accommodate thousands of independent variables, maintain accuracy even when a large proportion of data are missing (although we did not implement this missing data feature in our analysis), have been demonstrated to be among the most accurate statistical learning machines, and are capable of generating useful scaling of importance among predictors.

Results in this study were based on the average results of 10 forests of 500 trees each. Analyses were performed using the Random Forest as given in the R library (available at http://cran.us.r-project.org). Random Forest was originally developed by Breiman and Cutler (25).

RESULTS

Most of the 3,839 patients were young or middle-aged women (Table 1). The most common SLE-related manifestations were nephritis, thrombocytopenia, and psychosis, and the most common comorbid conditions were diabetes mellitus, urinary tract infection, pneumonia, chronic obstructive pulmonary disease, and congestive heart failure. A total of 109 patients (2.8%) died during the hospitalization.

In random forests developed using all the clinical and demographic predictors, the most important variables for predicting mortality were the Charlson Index, respiratory failure, the SLE Comorbidity Index, age, sepsis, nephritis, and thrombocytopenia (Table 2). The mean classification error rate, averaged over 10 forests of 500 trees each (a total of 5,000 trees), was 11.9%, demonstrating excellent predictive accuracy. A single classification tree demonstrating the relationships among the most important variables is shown in Figure 1.

Given that the Charlson Index and the SLE Comorbidity Index measure similar aspects of clinical severity, we also compared forests that selectively used either index. Predictive accuracy was significantly higher (and error rates were lower) in forests using the SLE Comorbidity Index than in forests using the Charlson Index (error rate 11.1% versus 13.2%; $P < 0.0001$ using 1-way analysis of variance).

To determine the importance of individual clinical manifestations, we repeated the analysis after excluding both the Charlson Index and the SLE Comorbidity Index. The most important predictors in these forests were respiratory failure, age, sepsis, thrombocytopenia, renal disease (either nephritis or chronic renal failure), congestive heart failure, and cerebrovascular accident (Table 2). Patient sex; ethnicity; insurance status; and other SLE manifestations, chronic comorbid conditions (e.g., diabetes mellitus), and acute complications (e.g., urinary tract infection or pulmonary embolus) were relatively much less important in correctly classifying patients. The mean classification error rate for these forests was lower than the mean classification error rate for those that included the Charlson Index and the SLE Comorbidity Index (10.4% versus 11.9%; $P < 0.0001$). A clustering algorithm, based on multidimensional scaling included as part of the random forests code, indicated that the patients who died did not comprise a subset unique in any particular combination of clinical or demographic features.

DISCUSSION

In this population-based study of hospitalized patients with SLE, in-hospital mortality could be accurately predicted using information on demographic characteristics and the presence or absence of common diagnoses. A

Table 1. Patient characteristics at the index hospitalization*

Characteristic	All patients (n = 3,839)	Patients who survived (n = 3,730)	Patients who died (n = 109)
Age, mean ± SD years	41.8 ± 15.6	41.7 ± 15.6	44.6 ± 17.0
Female sex	3,382 (88.1)	3,291 (88.2)	91 (83.5)
White	2,080 (54.2)	2,018 (54.1)	62 (56.9)
Black	843 (22.0)	823 (22.1)	20 (18.3)
Asian	486 (12.7)	474 (12.7)	12 (11.0)
Native American	20 (0.5)	20 (0.5)	0
Other ethnicity	370 (9.6)	355 (9.5)	15 (13.8)
Unknown ethnicity	40 (1.0)	40 (1.1)	0
Hispanic	1,028 (26.8)	998 (26.7)	30 (27.5)
Private insurance	1,627 (42.4)	1,581 (42.4)	46 (42.2)
Medicare	879 (22.9)	847 (22.7)	32 (29.4)
Public insurance	1,202 (31.3)	1,173 (31.4)	29 (26.6)
No insurance	131 (3.4)	129 (3.4)	2 (1.8)
Thrombocytopenia†	553 (14.4)	504 (13.5)	49 (45.0)
Hemolytic anemia	116 (3.0)	108 (2.9)	8 (7.3)
Pericarditis†	188 (4.9)	185 (4.9)	3 (2.8)
Pleuritis†	355 (9.2)	344 (9.2)	11 (10.1)
Nephritis‡	1,358 (35.4)	1,284 (34.4)	74 (67.9)
Renal failure‡	162 (4.2)	147 (3.9)	15 (13.8)
Seizures	306 (8.0)	283 (7.6)	23 (21.1)
Psychosis	383 (9.8)	365 (9.8)	18 (16.5)
Myocardial infarction‡	61 (1.6)	55 (1.5)	6 (5.5)
Unstable angina	13 (0.3)	13 (0.3)	0
Atrial fibrillation	117 (3.0)	111 (3.0)	6 (5.5)
Congestive heart failure‡	231 (6.0)	207 (5.5)	24 (22.0)
Nonspecific chest pain	42 (1.1)	42 (1.1)	0
Cerebrovascular accident‡	147 (3.8)	127 (3.4)	20 (18.3)
Hemiplegia§	10 (0.3)	9 (0.2)	1 (0.9)
Peripheral vascular disease‡	25 (0.6)	20 (0.5)	5 (4.6)
Chronic obstructive pulmonary disease§	286 (7.4)	278 (7.4)	8 (7.3)
Respiratory failure	94 (2.4)	52 (1.4)	42 (38.5)
Pulmonary embolus	23 (0.6)	20 (0.5)	3 (2.8)
Deep venous thrombosis	87 (2.3)	80 (2.1)	7 (6.4)
Peptic ulcer disease§	59 (1.5)	52 (1.4)	7 (6.4)
Gastritis	51 (1.3)	50 (1.3)	1 (0.9)
Mild liver disease§	45 (1.2)	38 (1.0)	7 (6.4)
Severe liver disease‡	18 (0.5)	15 (0.4)	3 (2.8)
Pancreatitis	51 (1.3)	45 (1.2)	6 (5.5)
Renal transplantation	39 (1.0)	39 (1.0)	0
AV fistula complications	73 (1.9)	71 (1.9)	2 (1.8)
Sepsis	93 (2.4)	65 (1.7)	28 (25.7)
Pneumonia	262 (6.8)	239 (6.4)	23 (21.1)
Urinary tract infection	430 (11.2)	412 (11.0)	18 (16.5)
Cellulitis	92 (2.4)	90 (2.4)	2 (1.8)
Osteomyelitis	8 (0.2)	8 (0.2)	0
AIDS‡	3 (0.1)	3 (0.08)	0
Diabetes mellitus‡	312 (8.1)	291 (7.8)	21 (19.3)
Cancer‡	27 (0.7)	27 (0.7)	0
Metastatic cancer‡	1 (0.03)	1 (0.03)	0
Dementia§	15 (0.4)	15 (0.4)	0
Dehydration	335 (8.7)	323 (8.6)	12 (11.0)
Syncope	18 (0.5)	17 (0.4)	1 (0.9)
Hip/pathologic fracture	25 (0.6)	19 (0.5)	6 (5.5)
Charlson Index			
0	1,783 (46.4)	1,774 (47.6)	9 (8.3)
1	487 (12.7)	472 (12.6)	15 (13.8)
2	1,183 (30.8)	1,154 (30.9)	29 (26.6)
3	301 (7.8)	259 (6.9)	42 (38.5)
≥4	85 (2.3)	71 (1.9)	14 (12.8)
SLE Comorbidty Index			
0	1,551 (40.4)	1,543 (41.4)	8 (7.3)
2	1,103 (28.7)	1,079 (28.9)	24 (22.0)
3	82 (2.1)	82 (2.2)	0
4	528 (13.8)	512 (13.7)	16 (14.7)
≥5	575 (15.0)	514 (13.8)	61 (56.0)

* All values are the number (percentage) unless otherwise indicated. AV = arteriovenous; AIDS = acquired immunodeficiency syndrome; SLE = systemic lupus erythematosus. Scores of 1 are not possible for the SLE Comorbidity Index.
† Included in the SLE Comorbidity Index.
‡ Included in the Charlson Index and the SLE Comorbidity Index.
§ Included in the Charlson Index.

Table 2. Relative importance of predictors of mortality during the index hospitalization*

Predictor	Forests including all predictors	Forests excluding the Charlson Index and SLE Comorbidity Index
Charlson Index	12.4	—
Respiratory failure	12.3	13.5
SLE Comorbidity Index	11.9	—
Age	8.5	9.3
Sepsis	5.7	6.5
Nephritis	3.8	4.9
Thrombocytopenia	3.8	6.2
Congestive heart failure	3.0	5.0
Nephritis or chronic renal failure	—	5.5
Cerebrovascular accident	2.1	3.9

* All other clinical and demographic predictors had relative importance values ≤3.0. SLE = systemic lupus erythematosus.

classification error rate of 10.4% was achieved using information that is collected routinely in clinical evaluations. Among the most important individual clinical manifestations that identified patients at increased risk of mortality were respiratory failure and sepsis, conditions known to be associated with a poor prognosis. Finding these conditions to be important predictors provides evidence of this approach's construct validity to classification. Additional important predictors of mortality were age, thrombocytopenia, nephritis or chronic renal failure, congestive heart failure, and cerebrovascular accidents.

Nephritis and thrombocytopenia have been noted as important predictors of survival in patients with SLE in many previous studies (2,3,5,7–9,11,13,14,16–18,22–24). In this study, these manifestations marked subsets of patients at increased risk for mortality. The association of thrombocytopenia or nephritis with mortality during the index hospitalization may be due to the effects of acute worsening of thrombocytopenia or nephritis at the time of the index hospitalization. Because laboratory results were not available in the data set, we could not examine this possibility. Similar effects may explain the association of congestive heart failure and cerebrovascular accidents with mortality, with acute exacerbations or acute events having greater importance in the risk of in-hospital mortality than a history of these conditions or a history of stable and compensated heart failure. The association of older age with the risk of mortality likely reflects the important contribution of age to the background risk of mortality.

Indexes summarize the information of several other variables, and might be expected to have better predictive ability than individual manifestations. When either the Charlson Index or the SLE Comorbidity Index was included as a potential classification variable, these indexes were by far the most important variables in identifying subsets of patients at high risk of mortality. However, classification error rates were somewhat higher when these indexes were included than when only individual manifestations were used as potential predictors. These results suggest that these indexes would be the preferred variables to use if only a single variable was to be chosen to classify risk of mortality, but that accuracy could be improved slightly using data on individual manifestations. When only 1 of these indexes was used, prediction was more accurate with the SLE Comorbidity Index than with the Charlson Index, possibly because the SLE Comorbidity Index captured risks specifically associated with manifestations of SLE better than the Charlson Index. These find-

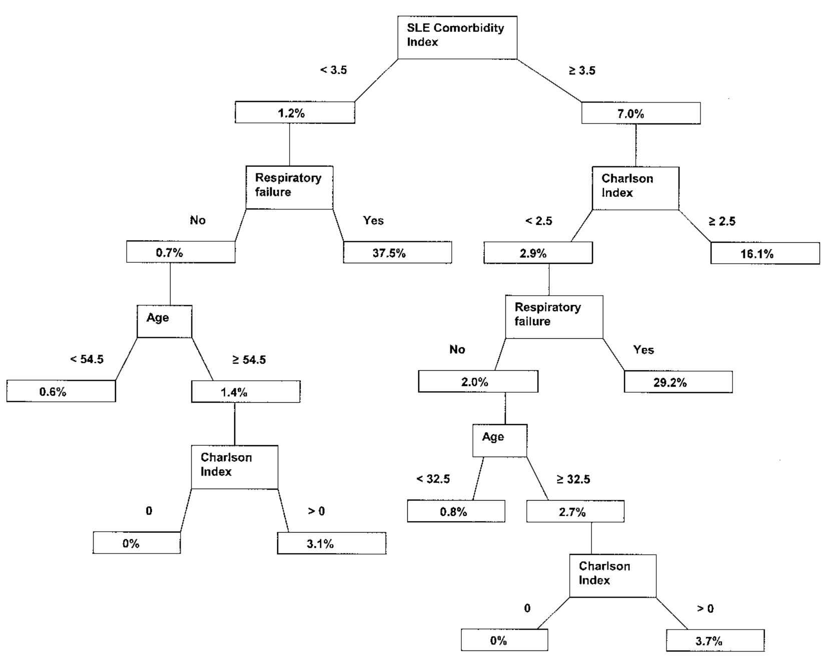

Figure 1. Single classification tree for the outcome of death during the index hospitalization, developed using all predictors. The variable at the top of the tree, the Systemic Lupus Erythematosus (SLE) Comorbidity Index, was selected by the program as the variable that resulted in the best separation of mortality risks between subgroups, with patients having an SLE Comorbidity Index score <3.5 having an estimated 1.2% risk of mortality and those with an SLE Comorbidity Index score ≥3.5 having a 7.0% mortality risk. These subgroups were then repeatedly split, based on the presence or absence of other clinical features, to achieve the final groups that were most homogenous in their mortality risks. This example uses a single matched draw balanced for the number of patients dead and alive. Mortality risks were those obtained when this tree was applied to the entire cohort.

ings support the use of the SLE Comorbidity Index when a single measure was to be used.

Our results are based on discharge diagnosis codes, which may not have been correct or complete for all patients. Undercoding is more common than overcoding in administrative databases, particularly for generally less urgent conditions such as hypertension or dementia (31,35,36). However, previous studies of these data have reported sensitivities >0.88 for diagnoses of cancer and chronic renal, liver, lung, and heart disease and specificities approaching 1.00 (31). Studies of other administrative databases have found the coding of idiopathic thrombocytopenic purpura to be highly accurate, and have found excellent agreement between Charlson scores derived from medical records and claims data (35–37).

The strengths of this study are the large population-based sample, the availability of information on multiple diagnoses, and the ability to identify rehospitalizations of individual patients, which allowed tracking of patients and their outcomes even in instances of interhospital transfers. This study also demonstrates the application of random forests as a useful tool to identify important classification variables using a procedure that provides cross-validated results and direct tests of predictive accuracy. The extensive cross-validation assures that the most important independent variables are identified. Random forests also provide protection from over-fitting the data, which can produce unreliable results. These properties help ensure that classification trees developed using the results of random forest analysis are true representations of the associations present in the sample of patients examined.

However, the study also has some limitations. Patients were identified by a physician diagnosis of SLE, and we could not confirm that all patients met the current classification criteria for SLE. However, for all patients, SLE was the principal diagnosis of the hospitalization, increasing our confidence that the diagnosis was likely correct for most patients. Because we limited the cohort to patients with a principal diagnosis of SLE, the predictors are specific to similar patients. Predictors may be different for patients hospitalized for reasons other than treatment of SLE. In addition, we did not have detailed clinical information on the degree of organ dysfunction, physiologic measures, or laboratory results, which could have influenced the predictors we identified. More accurate classification might have been possible using measures such as these, although the addition of clinical information has not been found to uniformly increase the accuracy of mortality prediction (38–42). In this study, classification of mortality was highly accurate even when the predictors were limited to only the presence or absence of a particular clinical manifestation. By using the outcome of in-hospital mortality, we missed deaths that occurred in the immediate postdischarge period but that might have been related to the hospitalization. However, we included data on mortality after interhospital transfers, which accounted for 11% of deaths observed and was likely the setting with the highest risk of death. In addition, the predictors identified here may not apply to mortality over a longer period.

Mortality risks in patients with SLE can be accurately predicted using clinical information from the hospitalization. Examination of detailed clinical information, including changes in laboratory measures, may provide important additional markers of mortality risks. Random forests provide a robust method to assess the accuracy and importance of these measures, and may be widely applicable to this and other classification problems.

REFERENCES

1. Trager J, Ward MM. Mortality and causes of death in systemic lupus erythematosus. Curr Opin Rheumatol 2001;13:345–51.
2. Wallace DJ, Podell T, Weiner J, Klinenberg JR, Forouzesh S, Dubois EL. Systemic lupus erythematosus: survival patterns: experience with 609 patients. JAMA 1981;245:934–8.
3. Ginzler EM, Diamond HS, Weiner M, Schlesinger M, Fries JF, Wasner C, et al. A multicenter study of outcome in systemic lupus erythematosus. I. Entry variables as predictors of prognosis. Arthritis Rheum 1982;25:601–11.
4. Halberg P, Alsbjorn B, Balslev JT, Lorenzen I, Gerstoft J, Ullman S, et al. Systemic lupus erythematosus: follow-up study of 148 patients. II. Predictive factors of importance for course and outcome. Clin Rheumatol 1987;6:22–6.
5. Reveille JD, Bartolucci A, Alarcon GS. Prognosis in systemic lupus erythematosus: negative impact of increasing age at onset, black race, and thrombocytopenia, as well as causes of death. Arthritis Rheum 1990;33:37–48.
6. Seleznick MJ, Fries JF. Variables associated with decreased survival in systemic lupus erythematosus. Semin Arthritis Rheum 1991;21:73–80.
7. Drenkard C, Villa AR, Alarcon-Segovia D, Perez-Vazquez ME. Influence of the antiphospholipid syndrome in the survival of patients with systemic lupus erythematosus. J Rheumatol 1994;21:1067–72.
8. Massardo L, Martinez ME, Jacobelli S, Villarroel L, Rosenberg H, Rivero S. Survival of Chilean patients with systemic lupus erythematosus. Semin Arthritis Rheum 1994;24:1–11.
9. Abu-Shakra M, Urowitz MB, Gladman DD, Gough J. Mortality studies in systemic lupus erythematosus: results from a single center. II. Predictor variables for mortality. J Rheumatol 1995; 22:1265–70.
10. Ward MM, Pyun E, Studenski S. Long-term survival in systemic lupus erythematosus: patient characteristics associated with poorer outcomes. Arthritis Rheum 1995;38:274–83.
11. Ward MM, Pyun E, Studenski S. Mortality risks associated with specific clinical manifestations of systemic lupus erythematosus. Arch Intern Med 1996;156:1337–44.
12. Stoll T, Seifert B, Isenberg DA. SLICC/ACR Damage Index is valid, and renal and pulmonary organ scores are predictors of severe outcome in patients with systemic lupus erythematosus. Br J Rheumatol 1996;35:248–54.
13. Jacobsen S, Petersen J, Ullman S, Junker P, Voss A, Rasmussen JM, et al. A multicentre study of 513 Danish patients with systemic lupus erythematosus. II. Disease mortality and clinical factors of prognostic value. Clin Rheumatol 1998;17:478–84.
14. Blanco FJ, Gomez-Reino JJ, de la Mata J, Corrales A, Rodriguez-Valverde V, Rosas JC, et al. Survival analysis of 306 European Spanish patients with systemic lupus erythematosus. Lupus 1998;7:159–63.
15. Kiss E, Regeczy N, Szegedi G. Systemic lupus erythematosus survival in Hungary: results from a single centre. Clin Exp Rheumatol 1999;17:171–7.
16. Bellomio V, Spindler A, Lucero E, Berman A, Santana M, Moreno C, et al. Systemic lupus erythematosus: mortality and survival in Argentina: a multicenter study. Lupus 2000;9: 377–81.
17. Mok CC, Lee KW, Ho CT, Lau CS, Wong RW. A prospective study of survival and prognostic indicators of systemic lupus erythematosus in a southern Chinese population. Rheumatology (Oxford) 2000;39:399–406.

18. Cook RJ, Gladman DD, Pericak D, Urowitz MB. Prediction of short term mortality in systemic lupus erythematosus with time dependent measures of disease activity. J Rheumatol 2000;27:1892–5.
19. Alarcon GS, McGwin G Jr, Bastian HM, Roseman J, Lisse J, Fessler BJ, et al, and the LUMINA Study Group. Systemic lupus erythematosus in three ethnic groups. VIII. Predictors of early mortality in the LUMINA cohort. Arthritis Rheum 2001;45:191–202.
20. Rahman P, Gladman DD, Urowitz MB, Hallett D, Tam LS. Early damage as measured by the SLICC/ACR damage index is a predictor of mortality in systemic lupus erythematosus. Lupus 2001;10:93–6.
21. Nived O, Jonsen A, Bengtsson AA, Bengtsson C, Sturfelt G. High predictive value of the Systemic Lupus International Collaborating Clinics/American College of Rheumatology damage index for survival in systemic lupus erythematosus. J Rheumatol 2002;29:1398–400.
22. Manger K, Manger B, Repp R, Geisselbrecht M, Geiger A, Pfahlberg A, et al. Definition of risk factors for death, end stage renal disease, and thromboembolic events in a monocentric cohort of 338 patients with systemic lupus erythematosus. Ann Rheum Dis 2002;61:1065–70.
23. Cervera R, Khamashta MA, Font J, Sebastiani GD, Gil A, Lavilla P, et al. Morbidity and mortality in systemic lupus erythematosus during a 10-year period: a comparison of early and late manifestations in a cohort of 1,000 patients. Medicine (Baltimore) 2003;82:299–308.
24. Bujan S, Ordi-Ros J, Paredes J, Mauri M, Matas L, Cortes J, et al. Contribution of the initial features of systemic lupus erythematosus to the clinical evolution and survival of a cohort of Mediterranean patients. Ann Rheum Dis 2003;62:859–65.
25. Breiman L, Cutler A. Random Forests. March 5, 2004. URL: http://www.stat.berkeley.edu/users/breiman/RandomForests/.
26. HCIA. The International Classification of Diseases, Ninth Revision, Clinical Modification: ICD-9-CM. Tenth ed. Baltimore: HCIA; 1993.
27. Zach A. New way to edit: discharge data review. Sacramento (CA): Office of Statewide Health Planning and Development; 1990.
28. Office of Statewide Health Planning and Development. Editing criteria handbook. Sacramento (CA): Office of Statewide Health Planning and Development; 1995.
29. Office of Statewide Health Planning and Development. Report of results from the OSHPD reabstracting project: an evaluation of the reliability of selected patient discharge data, July through December 1988. Sacramento (CA): Office of Statewide Health Planning and Development; 1990.
30. Office of Statewide Health Planning and Development. Second report of the California Hospital Outcomes Project: acute myocardial infarction. In: Volume Two: technical appendix. Sacramento (CA): Office of Statewide Health Planning and Development; 1996.
31. Romano PS, Mark DH. Bias in the coding of hospital discharge data and its implications for quality assessment. Med Care 1994;32:81–90.
32. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J Clin Epidemiol 1992;45:613–9.
33. Ward MM. Development and testing of a systemic lupus-specific risk adjustment index for in-hospital mortality. J Rheumatol 2000;27:1408–13.
34. Thorburn CM, Ward MM. Hospitalizations for coronary artery disease among patients with systemic lupus erythematosus. Arthritis Rheum 2003;48:2519–23.
35. Quan H, Parsons GA, Ghali WA. Validity of information on comorbidity derived from ICD-9-CCM administrative data. Med Care 2002;40:675–85.
36. Wilchesky M, Tamblyn RM, Huang A. Validation of diagnostic codes within medical services claims. J Clin Epidemiol 2004;57:131–41.
37. Segal JB, Powe NR. Accuracy of identification of patients with immune thrombocytopenic purpura through administrative records: a data validation study. Am J Hematol 2004;75:12–7.
38. Roos LL, Sharp SM, Cohen MM, Wajda A. Risk adjustment in claims-based research: the search for efficient approaches. J Clin Epidemiol 1989;42:1193–206.
39. Alemi F, Rice J, Hankins R. Predicting in-hospital survival of myocardial infarction: a comparative study of various severity measures. Med Care 1990;28:762–75.
40. Smith DW, Pine M, Bailey RC, Jones B, Brewster A, Krakauer H. Using clinical variables to estimate the risk of patient mortality. Med Care 1991;29:1108–29.
41. Hannan EL, Kilburn H Jr, Lindsey ML, Lewis R. Clinical versus administrative data bases for CABG surgery: does it matter? Med Care 1992;30:892–907.
42. Pine M, Norusis M, Jones B, Rosenthal GE. Predictions of hospital mortality rates: a comparison of data sources. Ann Intern Med 1997;126:347–54.